The location of the first AUG codons in cowpea mosaic virus RNAs

George P. Lomonossoff*, Michael Shanks*, Hans D. Matthes†, Mohinder Singh† and Michael J. Gait†

Department of Virus Research, John Innes Institute, Colney Lane, Norwich NR4 7UH, and Laboratory of Molecular Biology, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH, UK

Received 1 July 1982; Accepted 23 July 1982

ABSTRACT

We have made use of the known sequence of the 5' ends of both CPMV RNAs to synthesise an oligodeoxynucleotide which can prime second-strand DNA synthesis on full-length cDNA copies of both RNAs. By priming synthesis in the presence of dideoxynucleoside triphosphates, we have determined the positions of the first AUG codons in each RNA. These occur at positions 115 and 207 on M and B RNA respectively. By using a cloned double-stranded DNA fragment derived from near the 5' end of M RNA as a primer additional sequence from the 5' terminal region of M RNA has been obtained.

INTRODUCTION

The genome of cowpea mosaic virus (CPMV) consists of two RNA molecules of positive (messenger) polarity. Both RNA molecules have their 5' termini covalently linked to a small protein (Vpg) (1,2) and both are polyadenylated at their 3' ends (3). The 5' and 3' terminal sequences of both RNAs have been determined (4-6), revealing considerable sequence homology between the two RNAs in these regions, though overall the sequences are thought to be dissimilar (7). Both RNAs direct the synthesis of high molecular weight polypeptides which are subsequently cleaved to functional virus proteins (8,9). The primary product from bottom component (B) RNA is a single polypeptide of Mr 200,000, while middle component (M) RNA appears to code for two primary proteins (Mr's 105,000 and 95,000) of overlapping sequence (9). Surprisingly, peptide mapping data (9) suggest that the two M-coded proteins share both common N and C termini. A duplication in the RNA sequence near the initiation site for protein synthesis has been proposed to account for this finding (9). However the origins of protein synthesis on either RNA are unknown since no initiator AUG
codons were found in the 5' terminal 89 and 81 nucleotides of M and B RNA which have been sequenced so far (4,5).

We have made use of the known similarity in sequence at the 5' ends of M and B component RNAs (4,5) to synthesise an oligodeoxyribonucleotide corresponding to nucleotides 16-36 from the 5' end of both RNAs. We have used this oligonucleotide to prime second-strand DNA synthesis on full-length reverse transcripts of both RNAs in the presence of dideoxynucleoside triphosphates to obtain additional sequence at the 5' ends of the RNAs. This has enabled us to identify the positions of the first AUG codons on each RNA. Further sequence near the 5' end of M RNA was obtained by cloning Mbo 1 fragments of double-stranded cDNA in M13 and using one of the cloned fragments so obtained as a primer for further dideoxy sequencing on M cDNA.

MATERIALS AND METHODS

Oligonucleotide synthesis. The 21 base long oligodeoxyribonucleotide, d(A-G-G-T-T-T-G-A-T-A-A-A-A-G-C-G-A-A-C-G), was chemically synthesised by the solid phase phosphotriester method on a polynucleotide support as previously described (10), and purified by ion exchange and reversed phase high performance liquid chromatography (10). Based on 14 pmol of deoxynucleoside attached to the support, 0.238 pmol (60 A260 units) of the 21-mer was obtained in pure form (1.7% yield overall). The sequence was verified by 'wandering spot' analysis of kinase-labelled oligonucleotide (11).

cDNA synthesis. CPMV RNA was isolated from virus particles of the Nigerian isolate of CPMV strain SB by the method of Zimmern (12). Single-stranded cDNA was synthesised using an RNA concentration of 0.1 mg/ml and 0 dT12-18) (P-L Biochemicals) as primer (final concentration 0.01 mg/ml). Incubation was for 60 minutes at 42°C in 50 mM Tris-HCl (pH 8.3), 10 mM MgCl2, 1 mM DTT, 80 mM KCl, 0.5 mM deoxynucleoside triphosphates, including (α-32P) dCTP, 800 units/ml AMV reverse transcriptase (a gift of J.W. Beard) and 100-500 units/ml human placental ribonuclease inhibitor (Biotec, Madison). Synthesis was stopped by the addition of 2 volumes of 1:1 phenol/chloroform. Following deproteinisation the RNA strand was hydrolysed by the addition of EDTA to 15 mM and NaOH to 0.1 M and heat-
ing the mixture at 70°C for 15 minutes. The products were examin-
ed by electrophoresis on 1% alkaline agarose gels (13) and their
sizes determined using kinase-labelled denatured cauliflower mosa-
ic virus DNA fragments as size markers.

Full-length cDNA to M and B RNAs were purified by centrifuga-
tion through 15-30% sucrose gradients containing 1 M NaCl, 0.1 M
NaOH and 2 mM EDTA in a Beckman SW41 rotor for 23 hours at 35,000
rev/min and 15°C. Gradient fractions containing full-length M or
B cDNA were identified by electrophoresing a small portion on 1%
alkaline agarose gels (13). Appropriate fractions were neutrali-
sed with acetic acid and the DNA precipitated with ethanol.

Sequence determination at the 5' ends of the viral RNAs. Twenty
pmol of the synthetic 21-mer was phosphorylated with (γ-32P) ATP
using polynucleotide kinase. About 0.05 pmol of kinase-labelled
21-mer was annealed to approximately 0.2 µg of purified full-leng-
th M or B cDNA by boiling and slow cooling (14). The sequences
adjacent to the priming sites were determined by the chain termi-
nation technique (15) adapted for use with end-labelled primers
(16).

Cloning of DNA fragments. Double-stranded cDNA was synthesised by
annealing 4 pmol of 21-mer to unfraccionated 1st strand cDNA tran-
scribed from 2 µg of mixed M and B RNA and extending the chains
with DNA polymerase I (large fragment). The reaction was carried
out in 10 mM Tris-HCl, pH 7.4, 10 mM MgCl2, 80 mM KCl and 1 mM DTT
using 1.5 units of DNA polymerase (Boehringer) and 0.5 mM dNTPs
(including (a-32P) dATP). Incubation was for 2 hours at 20°C in
a total volume of 40 µl. The reaction was terminated by the addi-
tion of 2 volumes of 1:1 phenol/chloroform and the reaction mix-
ture fractionated on a 15-30% sucrose gradient buffered in 50 mM
Tris-HCl, pH 7.4, 1 mM EDTA. Centrifugation was for 23 hours at
35,000 rev/min and 15°C in an SW41 rotor. The fastest sedimenting
material was pooled and precipitated with ethanol. The double-
stranded DNA was digested with Mbo I (Biolabs) and the fragments
ligated into the Bam HI site of M13 mp2 Bam (17). Recombinant
phage were identified and DNA extracted from the virus as descri-
bred in (18). Dideoxy sequencing was carried out using a synthetic
17-long deoxyribonucleotide as primer (10). The replicative form
of a clone containing the sequence between bases 138 and 192 from the 5' end of M RNA (M32) was isolated and the insert released from the vector by cleavage with Eco R1 and purified by polyacrylamide gel electrophoresis. The restriction fragment was treated with exonuclease III and used as a primer for dideoxy sequencing of full-length M cDNA.

RESULTS

cDNA synthesis. Fig. 1 shows a 1% alkaline agarose gel of cDNA transcribed from a mixture of M and B RNAs in the presence of various amounts of human placental ribonuclease inhibitor. Use of the ribonuclease inhibitor dramatically increases incorporation into higher molecular weight material and two bands can be seen. The two bands were identified as full-length reverse transcripts of M and B RNA by comparison with kinase-labelled cauliflower mosaic virus DNA on alkaline agarose gels. The sizes of the single-stranded cDNAs were estimated at 3617±101 and 6208±145 bases (mean of six determinations) for M and B respectively in reasonable agreement with the published sizes of the RNAs (19,20). The inhibitor was subsequently used at a final concentration of 500 units/ml. Full-length cDNA transcripts of M and B which were suitable for priming with the synthetic 21-mer could readily be separa-

Figure 1. 1% alkaline agarose gel of cDNA synthesised in the presence of various amounts of human placental ribonuclease inhibitor. The concentration of inhibitor is shown in units/ml x 10⁻² at the top of the gel. Full-length cDNAs to M and B RNAs are marked with arrows. The broken arrow indicates the top of the gel.
Figure 2 (A). Nucleotide sequence at the 5' ends of M and B RNAs determined by priming on full-length cDNA with the synthetic 21-mer. The limits of previously determined sequence are marked with arrows.

(B). Additional sequence of M RNA determined by priming with the cloned Mbo 1 fragment isolated from M13 derivative M32. Samples were run for 1.5 or 3.5 hours on a 6% polyacrylamide gel. The position of the GATC at the end of the primer is marked by an arrow.
rated on 15-30% alkaline sucrose gradients.

Sequence at the 5' ends of the viral RNAs. Consistent with previous findings (16) we found it necessary to use kinase-labelled primer for sequencing with the synthetic 21-mer. If (α-32P) dATP was used very high backgrounds and spurious bands resulted for reasons that are not clear. Fig. 2 shows sequencing gels obtained by priming on either M or B cDNA with the 21-mer. The sequence could be read from position 42 from the 5' end of each RNA. The sequences near the priming sites were identical to those expected from previously published data (5), except that at position 68 on the M-RNA we find a C rather than a U. This change probably reflects the different isolates of the SB strain (Nigerian as opposed to Bi1) used for the sequencing studies since one base difference between the two isolates has already been found at position 10 of M RNA (4,5). The identity of the 5' sequences confirms that the 21-mer primes at the expected sites on the cDNAs. From the gels shown in Fig. 2A and others run for longer times the 5' sequences could be read through the previously known limits (nucleotides 89 and 81 for M and B RNAs respectively (5)) to nucleotide 195 of M RNA (Fig. 3). The first AUG codons occur at positions 115 and 207 on the M and B RNAs respectively. The sequence of the B RNA contains 3 ambiguities all of which involve C/U uncertainty (positions 92, 148 and 170). These ambiguities could be due either to sequencing artefacts or could reflect sequence heterogeneity in the leader sequence.

The sequences of M and B RNA preceding the first AUG codons have a number of common features. Both are U-rich (34.2% for M, 35.9% for B) and very G-poor (12.3% for M, 13.1% for B) and both sequences have alternating C-U stretches approximately 30 to 60 bases before the AUG codons (Fig. 3). In addition it was shown previously that the extreme 5' ends of the two RNAs show marked sequence homology (4,5) and these similarities presumably reflect the roles the two leader sequences have in common.

Extension of the M-RNA sequence. Two of the M13 clones produced as described earlier contained fragments derived from near the 5' end of M RNA (Fig. 3). One of the clones (M5) contained an insert corresponding to nucleotides 121-192 and contained an internal Mbo
Figure 3. Nucleotide sequences at the 5′ ends of M and B CPMV RNAs. The dotted line marks the sequence corresponding to the synthetic 21-mer and △ indicates the limits of previously determined sequence. Y indicates a pyrimidine residue. Mbo I sites (GAUC) are marked by square brackets. The alternating C-U sequences are also marked. The sequence of nucleotides 1-41 of both RNAs have been taken from previously published work (4,5), the base at position 10 on M RNA being shown as the two alternatives found. In the isolate used for these studies it is most probably an A (4).
This clone could have arisen through the cloning of a partial Mbo 1 digestion product or by fortuitous ligation of two neighbouring Mbo 1 fragments in the correct orientation. The other clone (M32) contained sequence between nucleotides 138 and 192. The sequence obtained from these clones confirmed the sequence already obtained using the 21-mer primer. Clone M32 was used to extend the known sequence of M-RNA. The insert was cleaved from the replicative form of the viral DNA by digestion with Eco R1 which cleaves at two sites flanking the Bam H1 site (17). The fragment was treated with exonuclease III and used to prime on M cDNA for dideoxy sequencing. The sequence that can be read from the gel shown in Fig. 2B extends from the GAUC at position 169 (Fig. 3) to nucleotide 414. The sequencing gel shown in Fig. 2B has one ambiguity. At position 201 (see Fig. 3) a band is visible in all four tracks though it is far more intense in the G track; the nucleotide at 201 has therefore been treated as a G. This extra sequence obtained at the 5' end of M RNA has enabled us to demonstrate that the AUG at position 115 is followed by an open reading frame sufficient to code for at least 100 amino acids. This provides some evidence that this AUG may indeed be the initiator of protein synthesis, especially since both the other reading frames contain termination triplets and have no AUGs (Fig. 3).

DISCUSSION

We have used the known sequence at the 5' ends of both CPMV RNAs to synthesise an oligodeoxynucleotide which can be used to prime second strand synthesis on full-length first strand cDNA transcribed from either M or B RNA. The yield of full-length first strand cDNA was greatly improved by the inclusion of human placental ribonuclease inhibitor in the reaction mixture. This has enabled sufficient material to be obtained for dideoxy sequence analysis of the 5' ends of both viral RNAs using the 21-mer as a primer.

The first AUG codons occur at positions 115 and 207 on M and B RNA respectively. At present it is not known whether these initiation codons are the functional initiation sites for the 105 and 200 kD proteins as no N-terminal sequence of these polypeptides is yet available. However in the case of M RNA sufficient seque-
nce has been obtained to show that this first AUG precedes an open reading frame of at least 100 amino acids. There is only one other AUG codon in the M RNA sequence and it is in-phase with the first one. Both the other reading frames contain three termination codons in the region following the first AUG. These two pieces of evidence, while circumstantial, suggest that protein synthesis does indeed start from the AUG at position 115. In the sequence following the first AUG there is no sign of the repeat proposed (9) to account the 105 and 95 kD proteins sharing both common N and C termini, though insufficient sequence information has been obtained to rule it out completely. The amino acid sequence coded for by the open reading frame is somewhat unusual, containing 9 cysteine and 9 serine residues in the hundred which have been determined. The evidence regarding whether the AUG at position 207 on B RNA functions as an initiator of protein synthesis even is less clear since less sequence data is available about the region following this AUG. As a rule protein synthesis starts at the first AUG codon on a messenger (21) making this AUG a likely candidate for the start of the Mr 200 protein. However poliovirus is a notable exception to this rule, initiation of translation occurring at the ninth AUG, 741 bases from the 5' end (22,23). This may be particularly relevant since poliovirus RNA is, like the two CPMV RNAs, initially translated into a large precursor polypeptide (24-26). In addition, poliovirus RNA has a similar structure to both CPMV RNAs, being both polyadenylated and linked to a Vpg (27,28). On the other hand the sequence immediately surrounding the AUG at position 207 on B RNA does correspond precisely to the sequence most favoured for eucaryotic initiation sites, AXXAUGG, (29) whereas this sequence is not found around the non-functional AUGs of poliovirus (22,23).

The RNA sequences preceding the first AUGs have several unusual features. As noted previously (4,5), the sequences of the two RNAs adjacent to the Vpg are very similar, there being only 5 differences in the first 44 bases. Both 'leader' sequences are rich in U and poor in G and such compositions are common for 5' leader sequences. The occurrence of runs of pyrimidines in the leader sequences of both RNAs means there are many opportunities for base-pairing with the purine-rich tract at the 3' end of euca-
ryotic 18S ribosomal RNA (30) which has been implicated in the in-
itiation of translation (31). A particularly striking feature of
the sequences is the occurrence of C-U alternating stretches.
These occur between positions 68 and 84 on M RNA (this work and
(5)) and between positions 149 and 161 on B RNA. Such alternating
stretches have not been found previously in the 5' leader sequen-
ces of eucaryotic messengers. In both cases these alternating se-
quenccs are followed by a stretch of sequence containing runs of
U residues (102-112 for M, 166-177 for B). The significance of
these features is unclear though the high U and low G content of
the 5' terminal sequence is similar to the situation at the 3'
ends of the molecules (6). The distances between the 5' termini
and the first AUG codons are somewhat greater than is normally
found for messenger RNAs but this does appear to be a common fea-
ture of plant virus genomic RNAs (32-35). In contrast the subge-
nomic messengers from TMV, BMV, CCMV, AlMV and TYMV all have rela-
tively short leader sequences (36-39), the length of the leader
presumably influencing the efficiency with which the messenger is
translated. At present it is unknown whether virion RNA itself
acts as a messenger in vivo, as it is possible that the actual
messengers may be derivatives, for instance lacking the Vpg, as is
the case for picornaviruses (40,41). Isolation of virus-specific
RNAs from polysomes may well resolve this point.

ACKNOWLEDGEMENTS

We thank J.R.O. Dawson for providing purified CPMV and A.O.
Jackson and J.W. Davies for their comments on the manuscript.

REFERENCES

1. Stanley, J., Rottier, P., Davies, J.W., Zabel, P. and van
J. Biochem. 92, 45-51.
206.
45-49.
309.
Acids Res. 7, 493-500.
<table>
<thead>
<tr>
<th>No.</th>
<th>Author(s)</th>
<th>Journal</th>
<th>Year</th>
<th>Volume</th>
<th>Pages</th>
</tr>
</thead>
</table>